Self-Supervised Representation Learning

Raul C. Incze

Contents

A little about myself...

Education

- BSc, Computer and Information Sciences - Babeș-Bolyai University (2015)
- MSc, Applied Computational Intelligence - Babeș-Bolyai University (2017)

Work Experience

- Game Developer (4 years)
- Machine Learning Engineer (3 years)

cognifeed

Cogni-what?

Accessible Machine Teaching:

- Empowers people with the domain knowledge to train models.
- Sample and label efficiency.

Cogni-what?

Accessible Machine Teaching:

- Empowers people with the domain knowledge to train models.
- **Sample and label efficiency.**
  - Creating new datasets for each task is expensive.
  - Some domains are supervision-starved.
  - There are many unlabelled data samples.

Sample Efficiency In Machine Learning

$$\frac{1}{\varepsilon(1-\sqrt{\varepsilon})}\left[2d\ln(6/\varepsilon)+\ln(2/\delta)\right]$$

- VC dimension ~ the effective number of parameters (expressiveness).
- In practice, the number of samples = 10 x VC dimension.[2]
- Deep nets => huge VC dimension

1. Bounding sample size with the VC dimension, Shawe-Taylor, J. et al, 1993
2. The VC Dimension - A measure of what it takes a model to learn, Yaser Abu-Mostafa, 2012

Conditioning is important

- Priors
- Regularization

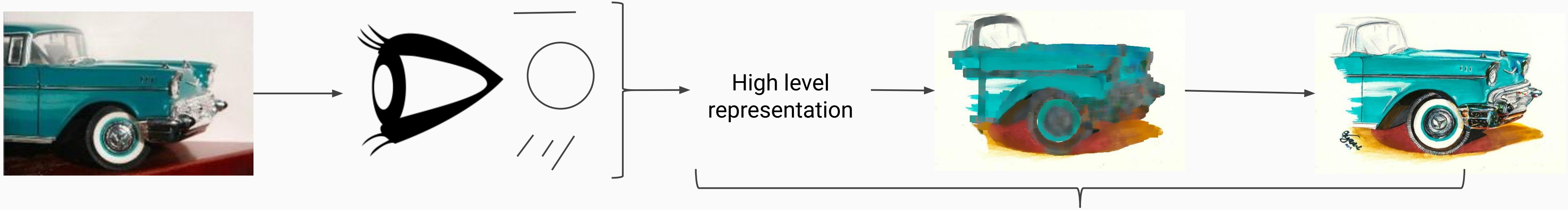
$$\tilde{O}\left((m+r)/\varepsilon^2\right)$$

1. The Model Complexity Myth, Jake VanderPlas, 2015
2. How Many Samples are Needed to Estimate a Convolutional Neural Network, Simon S.Du et al, 2018

We can do better...

Representations

The Human Brain and Representation Learning



Some takeaways:

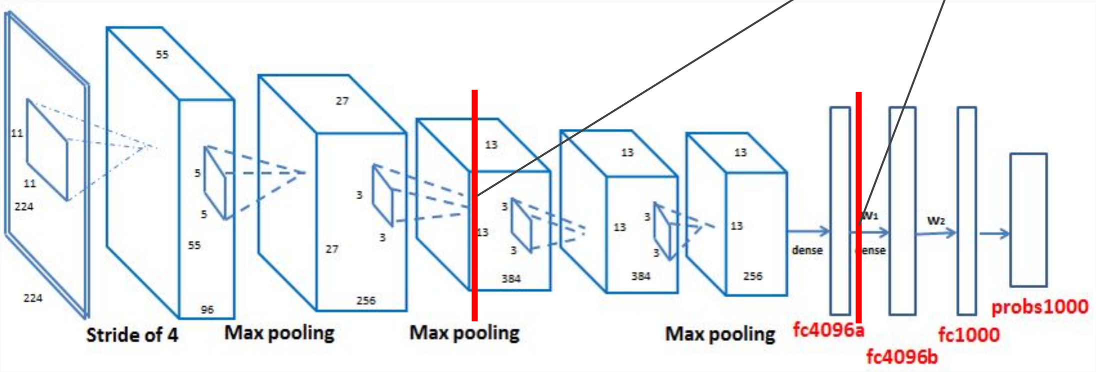
- We store representations, not snapshots.
- We use these representations to reason (predict, classify) new inputs and recall old ones.
- We don't need "supervision" to generate representations.

1. The Human Brain Recalls Visual Features in Reverse Order Than It Detects Them, Zuckerman Institute, 2017

Representations in Machine Learning

- Also known as “feature learning”.
- It’s a **consequence** of deep supervised learning.

Activations for input (logits) = features / representations

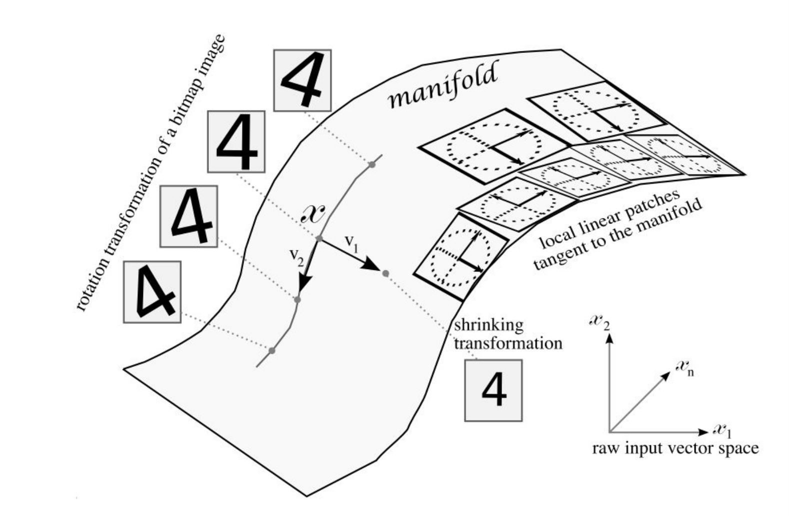


We can also make models learn representations **explicitly**.

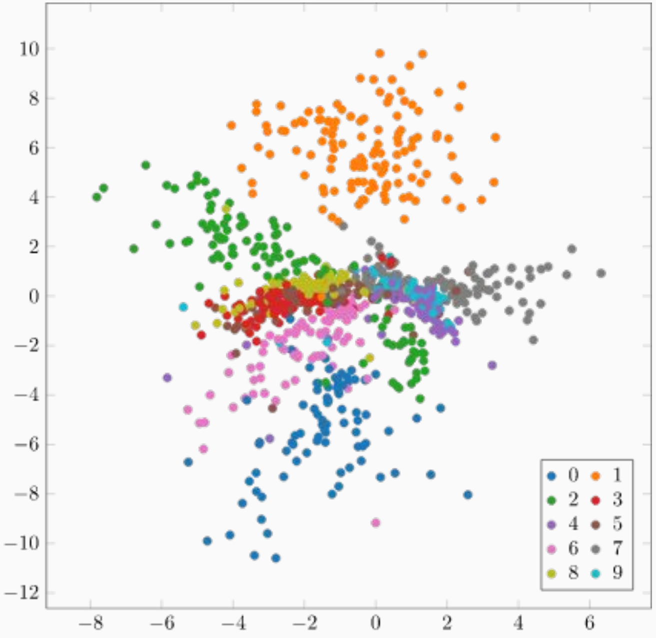
What makes a good representation?

- **Smoothness**
- Multiple explanatory factors
- Depth
- **Shared factors across tasks**
- **Manifolds**
- **Natural clustering**
- **Temporal and spatial coherence**
- Sparsity
- Simplicity of factor dependencies

1. Representation Learning: A Review and New Perspectives, Yoshua Bengio et. al, 2014



Smoothness



Natural clustering

Self-Supervision

Self-supervision?

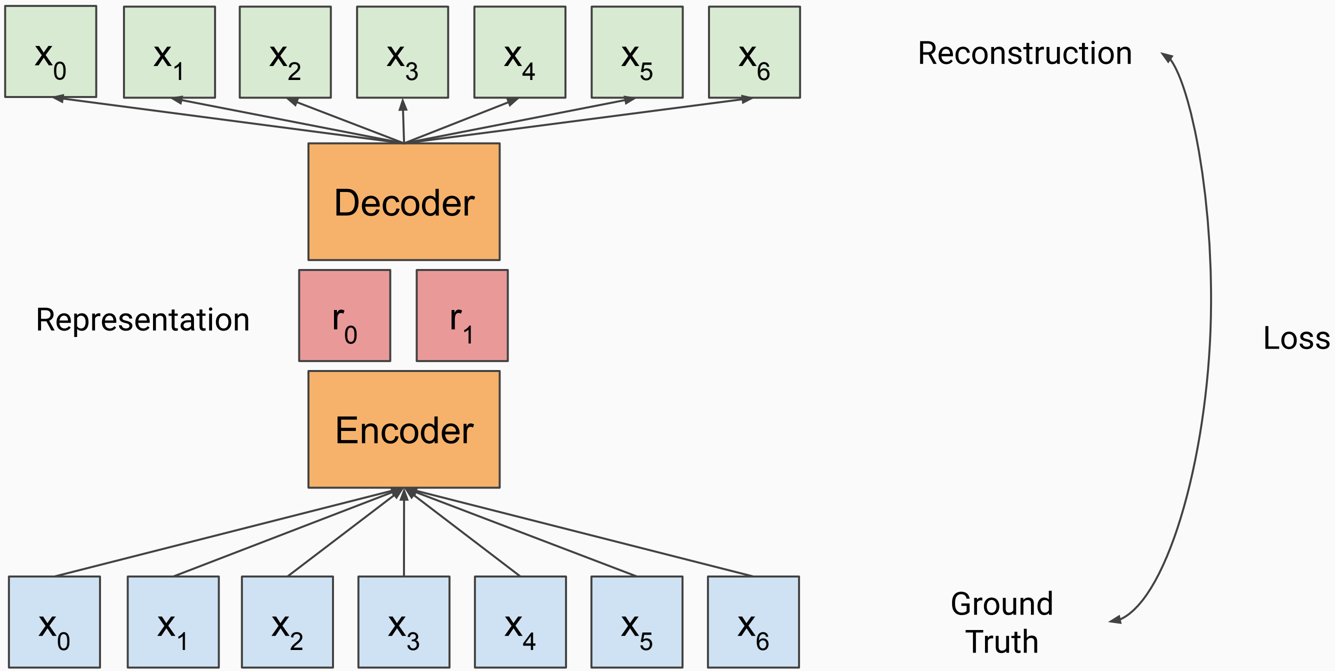
- A form of unsupervised learning.
- Data itself provides the supervision.

Two ways of achieving this:

- We withhold part of the input data and make our model predict it.
- Making up pretext tasks.

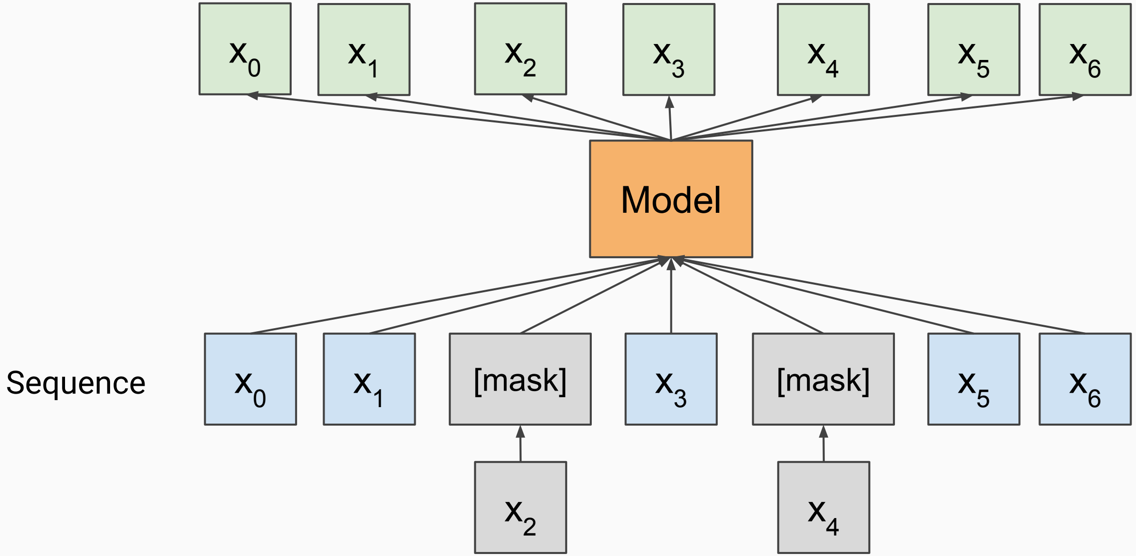
Withholding Data - Autoencoding (AE) Methods

Autoencoding (AE) Methods



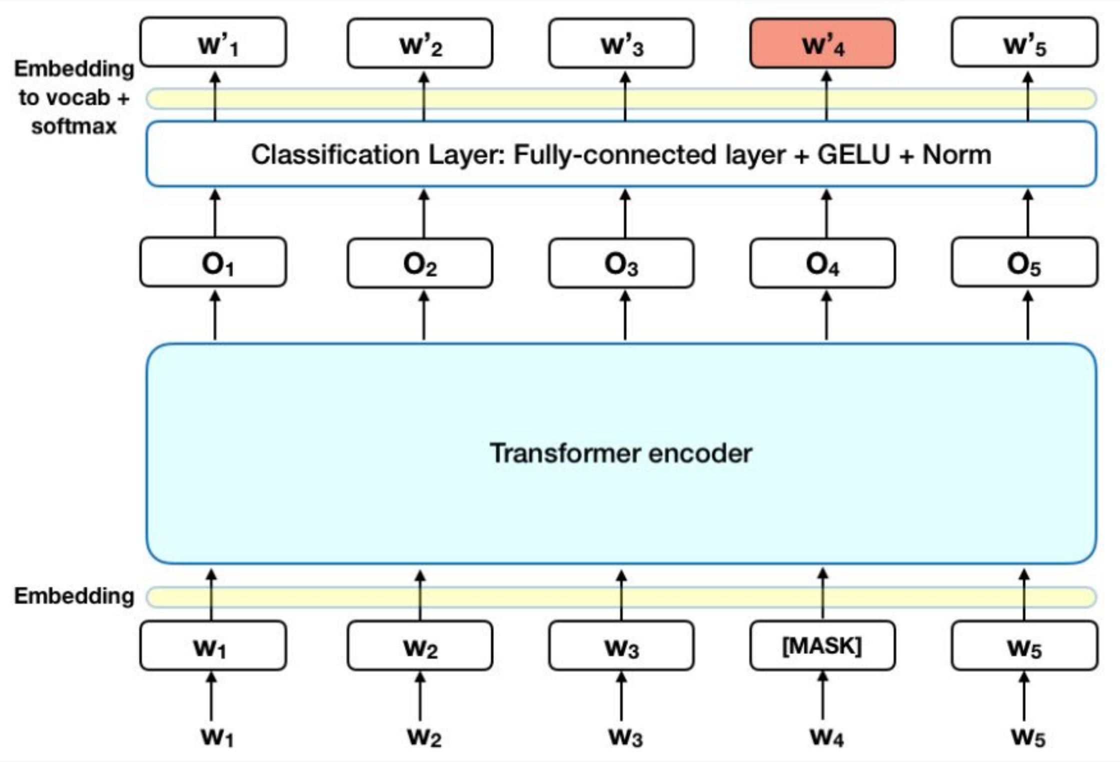
No information withheld (or arguably all the input is withheld from the decoder), representation is quite arbitrary, reconstruction can be poor.

Autoencoding (AE) Methods



- Denoising autoencoder (noise = input values set to 0/mask value).
- Works under the assumption that the choice for x_2 is independent from the choice of x_4.
- Natural Language breaks this assumption but this still works well on NLP tasks.
- Model = Transformer => BERT

Study Case: Bidirectional Encoder Representations from Transformers (BERT)



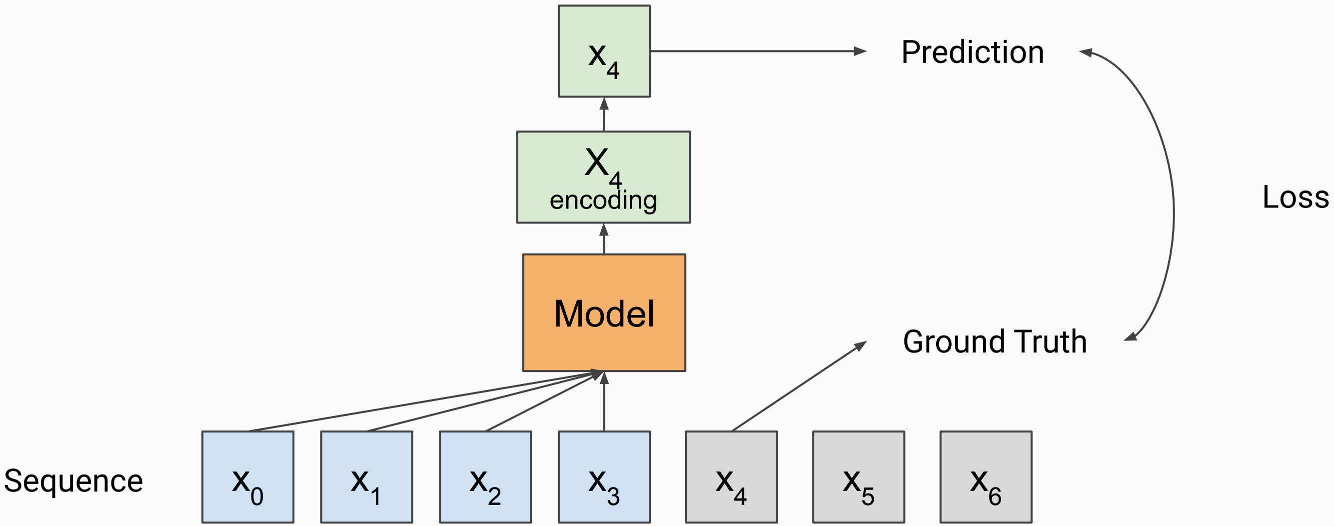
Fine-tuning approach	
$\text{BERT}_{\text{LARGE}}$	96.6
$\text{BERT}_{\text{BASE}}$	96.4
Feature-based approach ($\text{BERT}_{\text{BASE}}$)	
Embeddings	91.0
Second-to-Last Hidden	95.6
Last Hidden	94.9
Weighted Sum Last Four Hidden	95.9
Concat Last Four Hidden	96.1
Weighted Sum All 12 Layers	95.5

Results on Named Entity Recognition

1. Attention Is All You Need, Ashish Vaswani et. al, 2017
2. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Jacob Devlin et. al, 2018

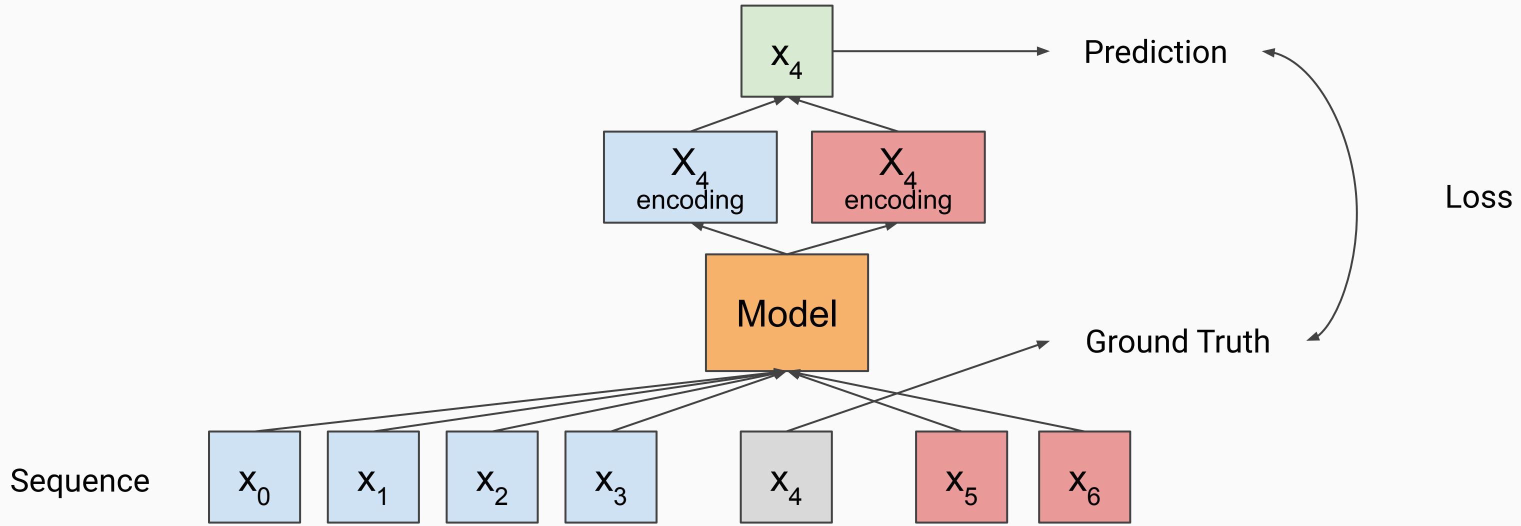
Withholding Data - Autoregressive (AR) Methods

Autoregressive Methods



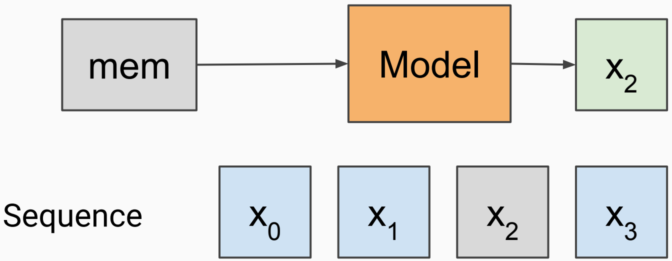
- Assumes that for the prediction of x_t all we need is x_0, ..., x_{t-1}
- Model = Transformer => GPT (OpenAI language model).

Autoregressive Methods

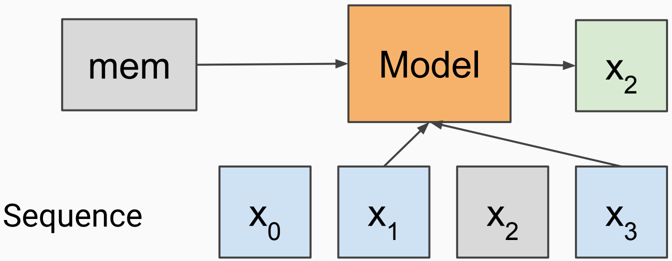


- Assumes the whole bidirectional context.
- Model = LSTM => ELMO.

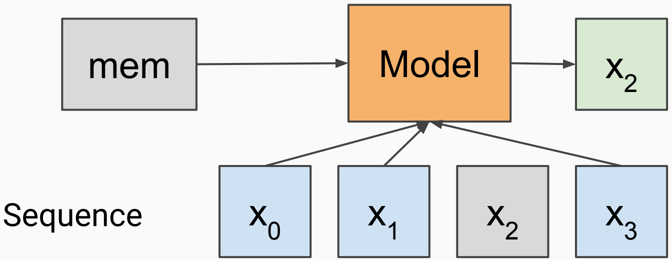
Autoregressive Methods



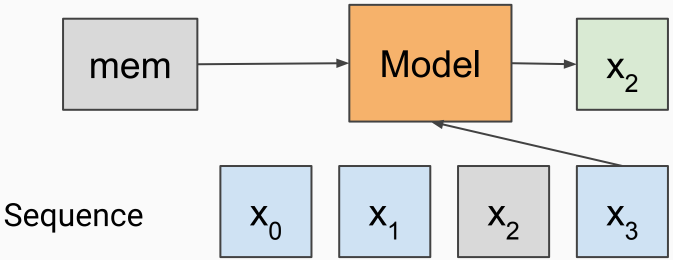
Factorization order: 2□1□3□0



Factorization order: 1□3□2□0



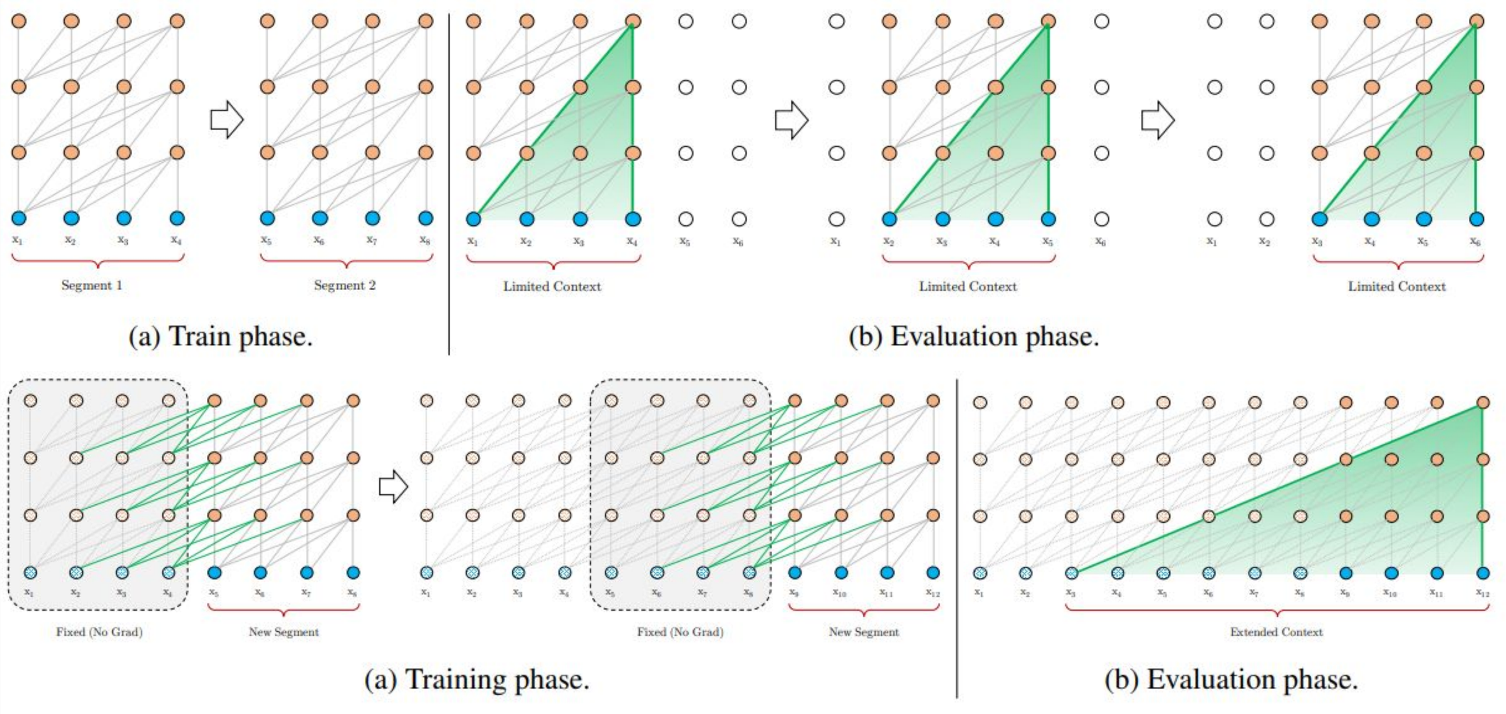
Factorization order: 0□3□1□2



Factorization order: 3□2□0□1

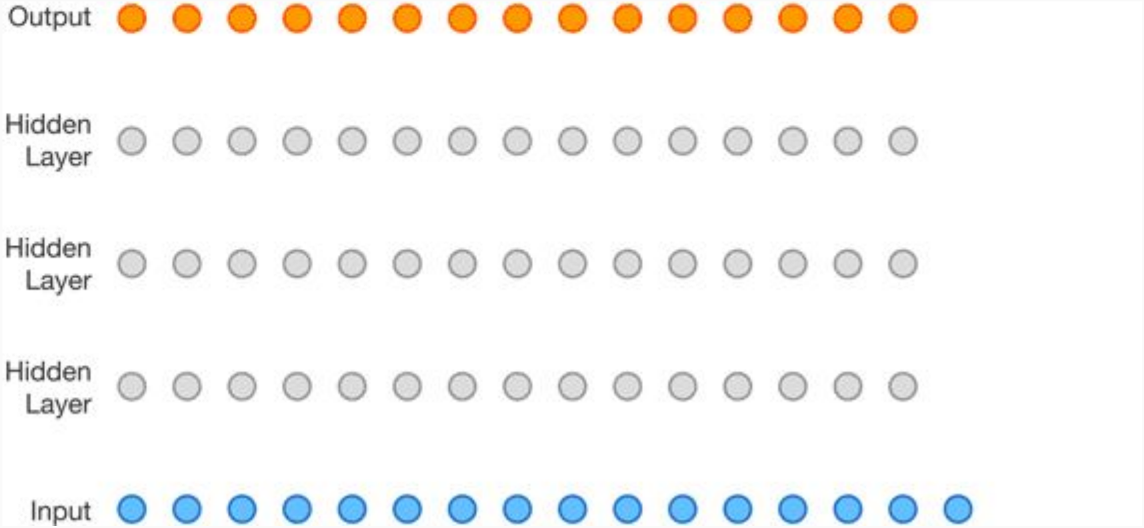
Case Study: XLNet

Factorization + Transformer-XL

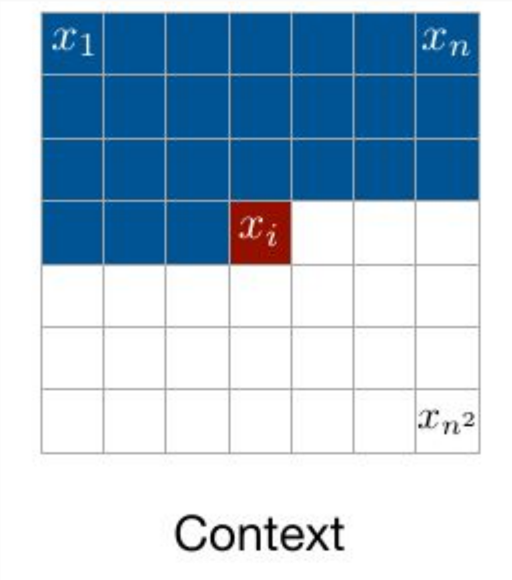


1. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context, Zihang Dai et. al, 2019
2. XLNet: Generalized Autoregressive Pretraining for Language Understanding, Zhilin Yang et. al, 2019

Other Autoregressive Models



WaveNet

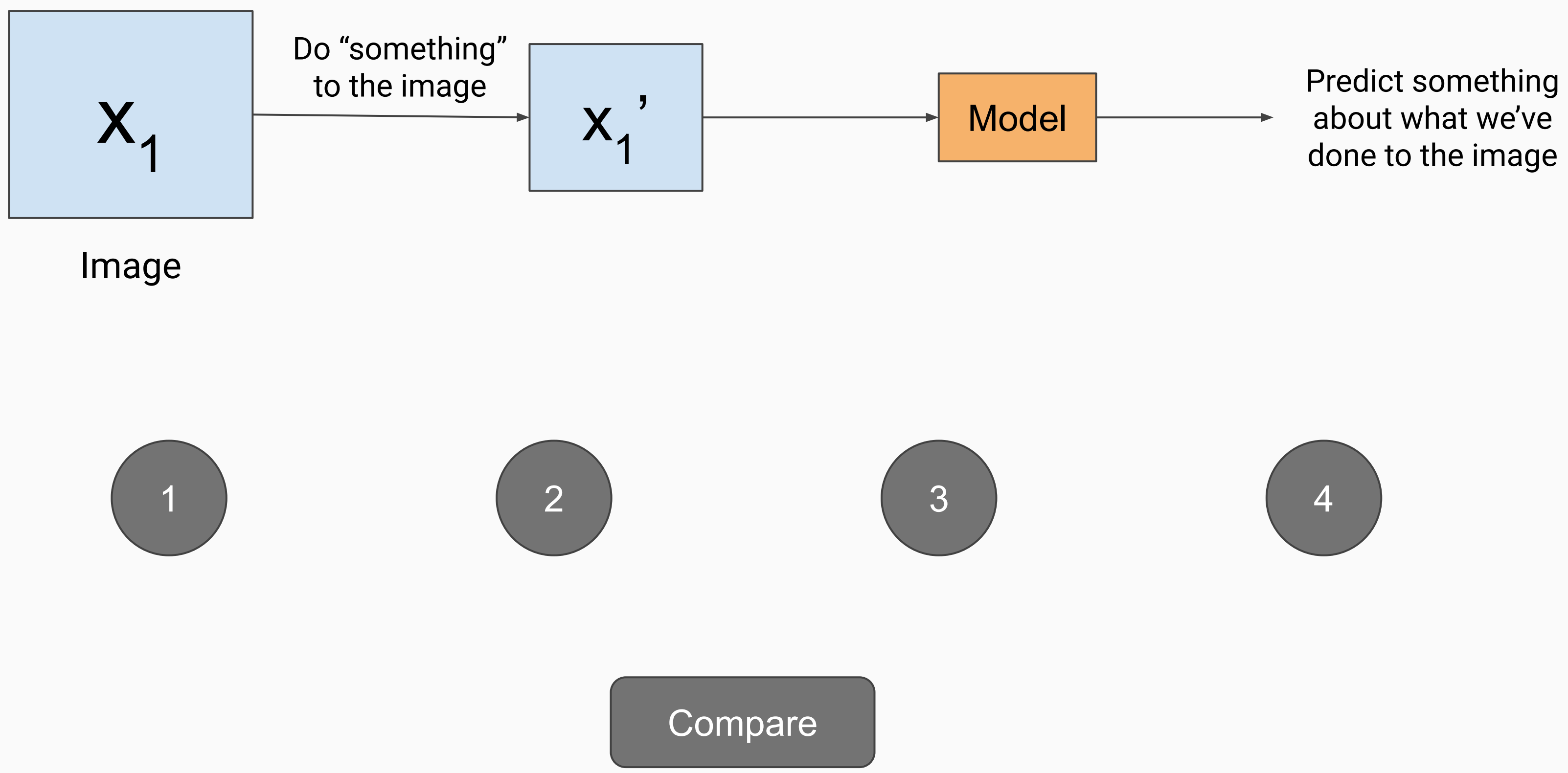


PxelRNN

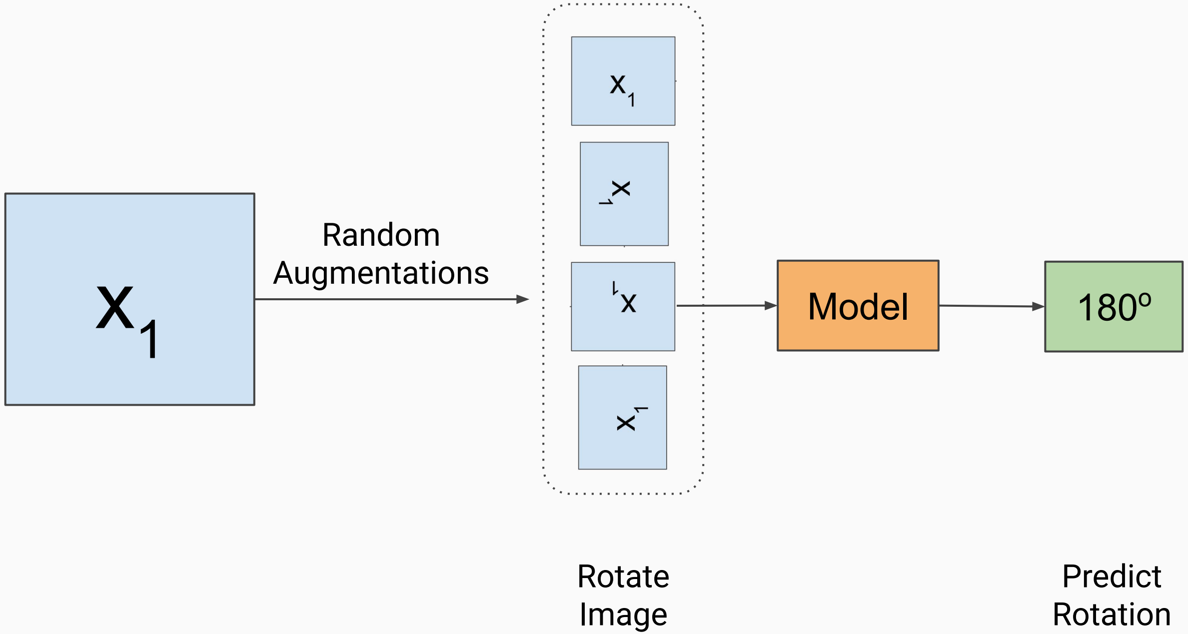
1. WaveNet: A Generative Model for Raw Audio, Aaron van den Oord et. al, 2016
2. Pixel Recurrent Neural Networks, Aaron van den Oord et. al, 201

Pretext Tasks

Brainstorming Exercise - Pretext Tasks For Visual Representation Learning

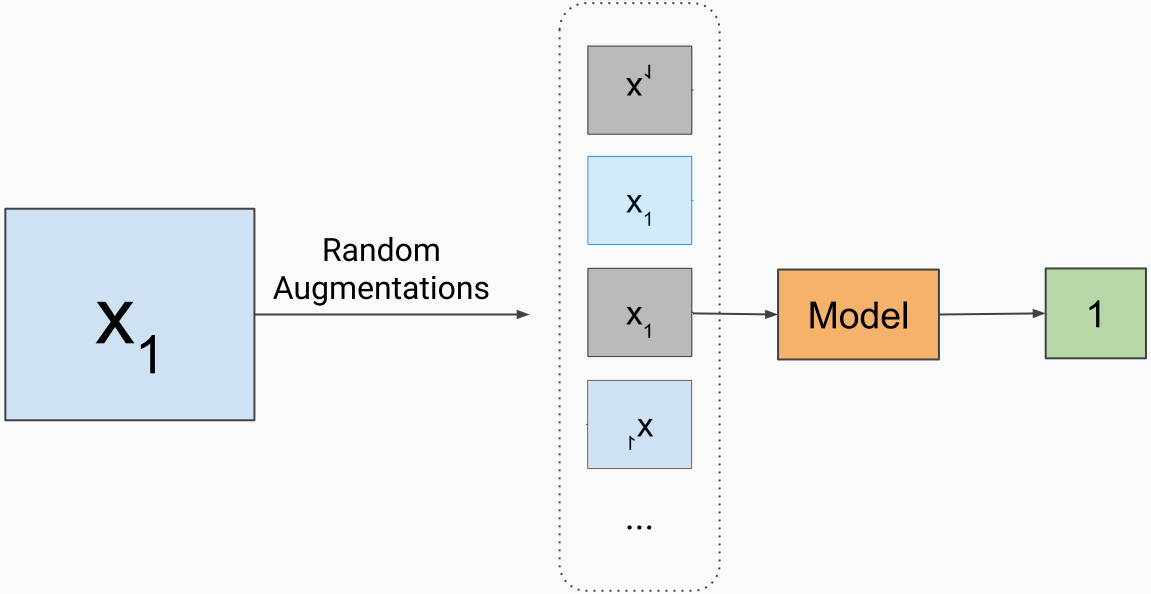


Rotation



Back

Exemplar



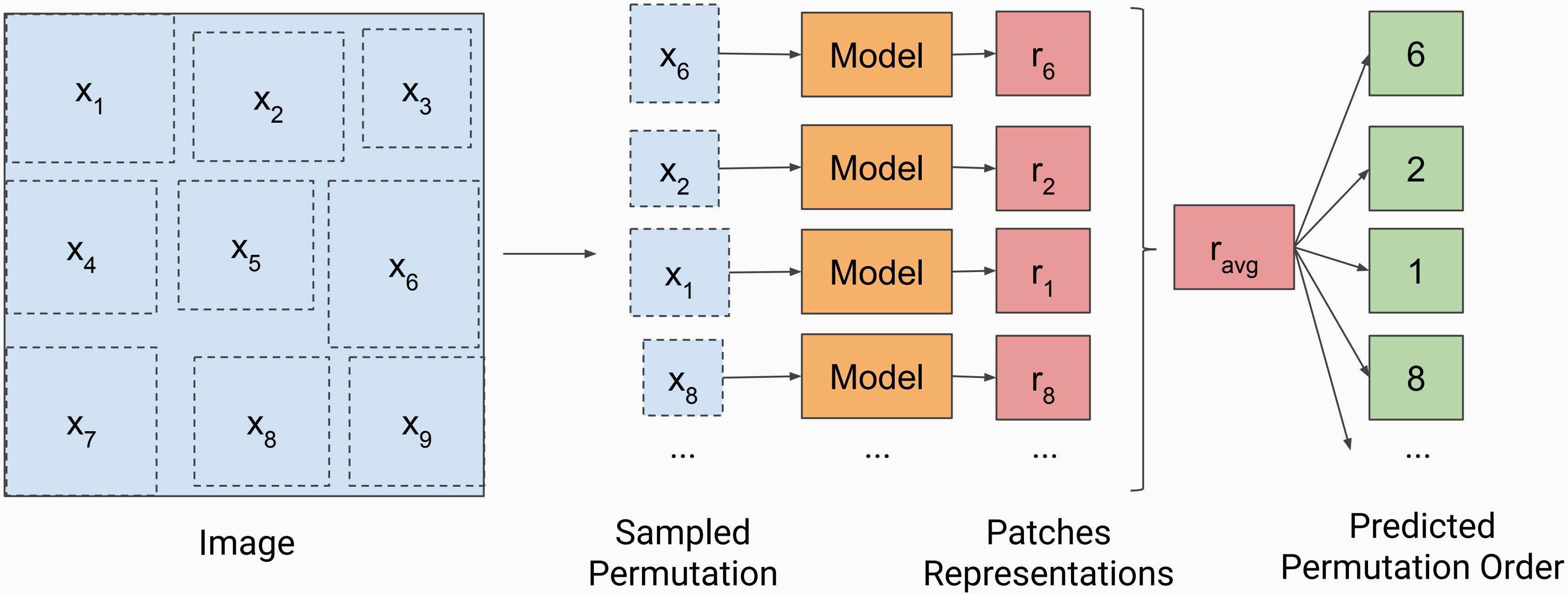
Every original image is an "exemplar" and has a corresponding class

Label all augmentations of x_1 with x_1's exemplar label "1"

Model has to predict x_1 exemplar label

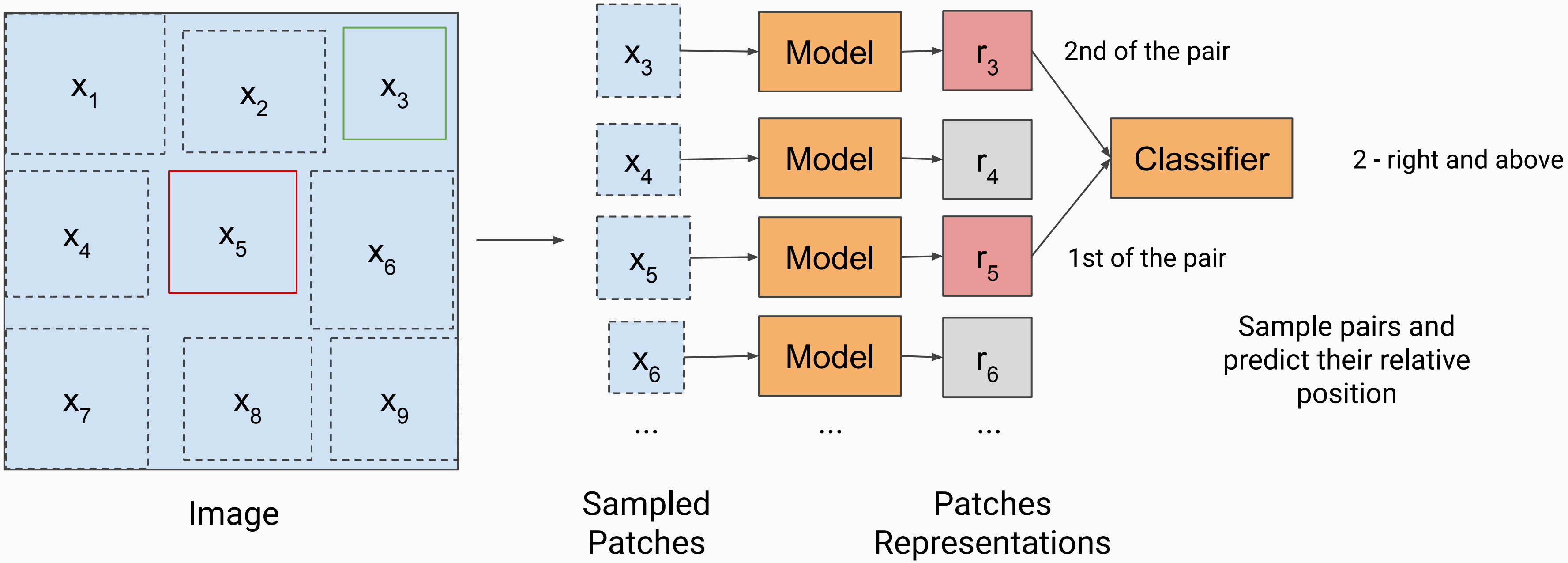
Back

Jigsaw


x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
x_9
Image
x_6
x_2
x_1
x_8
...
Sampled Permutation
Model
Model
Model
Model
...
r_6
r_2
r_1
r_8
...
Patches Representations
r_{avg}
6
2
1
8
...
Predicted Permutation Order

Back

Relative Patch Location



Sample pairs and predict their relative position

Back

Comparison

Model	Rotation				Exemplar			RelPatchLoc		Jigsaw	
	4×	8×	12×	16×	4×	8×	12×	4×	8×	4×	8×
RevNet50	**47.3**	**50.4**	**53.1**	**53.7**	**42.4**	45.6	46.4	40.6	45.0	40.1	43.7
ResNet50 v2	43.8	47.5	47.2	47.6	**43.0**	45.7	46.6	42.2	46.7	38.4	41.3
ResNet50 v1	41.7	43.4	43.3	43.2	**42.8**	**46.9**	**47.7**	**46.8**	**50.5**	**42.2**	**45.4**
RevNet50 (-)	45.2	**51.0**	**52.8**	**53.7**	38.0	42.6	44.3	33.8	43.5	36.1	41.5
ResNet50 v2 (-)	38.6	44.5	47.3	48.2	33.7	36.7	38.2	38.6	43.4	32.5	34.4
VGG19-BN	16.8	14.6	16.6	22.7	26.4	28.3	29.0	28.5	29.4	19.8	21.1

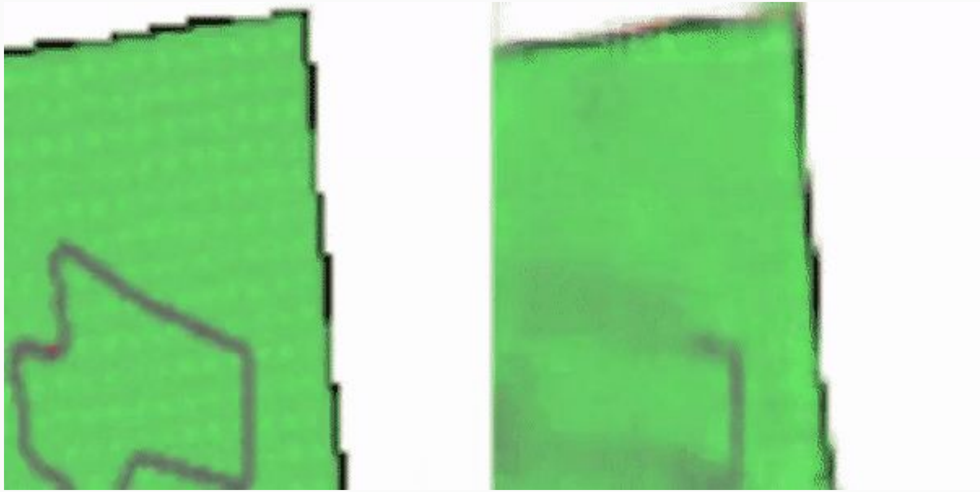
1. Revisiting Self-Supervised Visual Representation Learning, Alexander Kolesnikov et. al, 2019

Putting Representations to Good Use

Representations In Few-Labels Scenarios

- Pretrain on vast amounts of data in a self-supervised manner and then:
  - Fine-tune on known labels.
  - Use learned representations to train a shallow model.
- Use representations' properties (natural clustering) for label propagation.
- Active Learning
  - Representations are learned.
  - Estimator is trained on available labels.
  - Estimator predicts labels for unlabeled instances.
  - Forwards the most uncertain instances to the user to be corrected.
- Co-Training
  - Learn two different representations for each data point => two views.
  - Train an estimator on each view.
  - Each estimator predicts labels for unlabeled instances and adds the most confident prediction to the label pool of the other estimator.

Self-Supervised Representations in Reinforcement Learning



Agents learning in their own "dreams"[1]

(a) learn to explore on Level-1 (b) explore faster on Level-2

Curiosity-driven exploration[2]

1. World Models, David Ha and JÜRGEN SCHMIDHUBER, 2018
2. Curiosity-driven Exploration by Self-supervised Prediction, Deepak Pathak et. al, 2017

Question Time